参赛密码 ————————

（由组委会填写）

# “华为杯”第十三届全国研究生数学建模竞赛

| | |
|---|---|
| 学　校 | 上海交通大学 |
| 参赛队号 | **10248163** |
| 队员姓名 | **1.强思维** |
| | **2.王海洋** |
| | **3.李龙元** |

参赛密码 ————————

（由组委会填写）




# “华为杯”第十三届全国研究生数学建模竞赛

题 目 具有遗传性疾病和性状的遗传位点分析建模

## 摘 要：

本文针对位点的特征选择问题，运用假设检验、模型选择、回归随机森林、逻辑回归、岭回归、多标签随机森林等方法对问题进行求解。

对于问题 1，本文根据分析需求，采用了两种编码方式：(1)基于位点碱基分布的{0,1,2}编码方式。(2)独热编码(One-hot)编码方式，分别用于解决假设检验和分类回归问题。

对于问题 2，本文提出了一种高维数据保证质量的特征选择模型，模型融合了特征选择问题算法的过滤式选择模型和包裹式选择模型，首先基于信息增益对所有特征进行特征的相关性评价，选出相关性最显著的特征作为候选特征，最无关的特征作为“阴影”特征，然后针对筛选过的特征采用回归随机森林进行变量重要性评价，通过迭代的方式过滤重要性低于阴影特征四分位点的候选特征，选出真正重要的特征。经过结果检验，可以发现本模型选择出了最重要的 40 个位点，10 折交叉验证中对于遗传疾病 A 预测的准确率达到 72%。

对于问题 3，本文首先采用互信息模型衡量了位点之间相关性，验证了相同基因的不同位点之间存在连锁不平衡性质，因此不能假设位点之间相互独立，之后采用了逻辑回归和岭回归分类模型分别对各个基因中的可疑位点和疾病之间进行回归分析，并基于决定系数 $R^2$ 挑选出了 20 个与疾病最相关的基因，这些基因通常包含较多的可疑位点或其中某个位点的重要性评分较高。经过结果检验，验证了获取的最可能的这 20 个基因对于遗传疾病 A 预测的准确率达到 70%。

对于问题 4，本文根据多个位点对应多个性状这一特点本文把它归为多标签分类问题(Multi-label Classification)建模。首先对性状之间的相关性作了分析，结合数据本身特点以及性状之间的强相关性，本文选择算法适应法(Algorithm Adaptation)，以 Random Forest of ML-C4.5 (RFML-C4.5)建立模型。通过假设检验

与多性状投票相结合的方式进行特征筛选，通过 MLRF 模型选择 40 个最重要特征，通过 10 折交叉验证得到 72.1%的准确率，并通过对比试验证明了多标签模型较单标签模型的优势。

关键词　遗传统计学 全基因组关联性分析(GWAS) 位点(SNPs) 特征选择 回归分析 多标签分类

目录

# 1　问题背景与重述

## 1.1　问题背景

随着基因技术的发展，根据基因进行个性化医疗的需求越来越强烈。遗传位点的选择、分析、识别是全基因组分析中最重要的任务。在应用于人类全基因组数据的研究中，参与者分为病例组合对照组，其中病例组具有某种疾病或性状，而对照组不具有某种疾病或性状。以血压为例，每个参与者首先根据临床表现被分类，然后获取每个人的一段 DNA，从中提取遗传位点序列。如果某等位基因在患病群体中出现的更加频繁，那么就可以说其相关的遗传位点与人类所患有的疾病相关，在某种方式上影响了患该疾病的风险。对于全基因组遗传位点，传统研究方法主要是计算位点和遗传性状的相关度，通过假设检验的方式测量所有遗传位点的 p 值，可以过滤掉与疾病或性状不相关的遗传位点。

这个任务的挑战性在于 GWAS 相关数据具有非常高维度的特征，而且大部分的遗传位点与疾病或者性状无关。在全基因组分析中，先进的机器学习方法可以很好地识别与疾病或性状具有很大相关性的遗传位点。

## 1.2　问题重述

根据提供的位点、基因、病例数据，研究一下问题：

问题一：位点碱基对编码问题

每个样本，采用碱基（A，T，C，G）的编码方式标识每个位点的信息，每个位点在不同样本中，有三种不同的编码组合，根据分析需求，设计编码方式将每个位点的碱基（A，T，C，G）的字符编码方式转化为数字编码方式。

问题二：单个遗传疾病与位点的相关性分析问题

遗传疾病 A 可能与一个或多个位点有关联，可以对样本的健康状况和位点编码对比分析来确定致病位点。本题所需的数据包括 phenotype.txt 和 genotype.dat 文件。其中 phenotype.txt 文件中包括了 1000 个样本具有遗传疾病 A 的信息，一列 0 和 1 组成的数据{0,0,….,1,1}，其中 0 代表样本没有疾病 A，1 代表样本患有疾病 A。genotype.dat 文件包含了 1000 个样本在某条染色体片段上的 9445 个位点的编码信息，其中第 1 行是位点的名称，第 2-1001 行是 1000 个样本在 9445 个位点的编码信息。本题要求根据提供的数据，提供方法，找出与疾病 A 最相关的一个或者多个位点。

问题三：单个遗传疾病与基因的相关性分析问题

基因是 DNA 上具有遗传效应的片段，每个基因片段上包含了若干个位点。对于某疾病，可能与某条基因包含的全部或者部分位点相关，本题提供的 gene_info 中，包含了 300 个 dat 文件，分别表示 300 个基因的信息，每个文件

中，包含若干个位点的名称。本问要求根据提供的数据，分析遗传疾病 A 与每个基因的相关性，提供方法，找出与疾病 A 最相关的一个或多个位点。

问题四：多个相关遗传疫病或症状与位点的相关性分析问题

人的许多遗传疾病和性状具有明显的相关性，把相关的性状或疾病放在一起研究，有助于提高发现致病位点的能力。本体提供 mluti_phenos.txt 中，给出了 1000 个样本的 10 个相关症状的病例信息。本问要求根据提供的数据，找出与 10 个相关疾病或症状均有关联的位点。

# 2　模型假设

1. 假设样本中碱基对标注无误，数据真实可靠
2. 假设样本对应的个体相互独立，不存在地域差异和人群分层
3. 假设存在一个或多个和疾病相关的致病位点或基因
4. 假设样本量足够大，能够挖掘出相关致病位点或基

# 3 符号说明

| 符号 | 符号说明 |
| --- | --- |
| $\alpha_i$ | 第 i 个位点 |
| $G=\{\alpha_i \mid i=1,\ldots,m\}$ | 样本中所包含的所有位点的集合 |
| $\lambda_i$ | 第 i 个性状 |
| $L=\{\lambda_i \mid i=1,\ldots,n\}$ | 所包含的所有性状的集合 |
| $\mathbf{X}_i$ | 第 $i$ 个样本的特征(位点)向量 |
| $\mathbf{Y}_i$ | 第 i 个样本的特征标签(性状)集 |
| $D=\{\mathbf{X}_i,\mathbf{Y}_i \mid i=1,\ldots,m\}$ | 样本集合 |
| $S$ （$S\subseteq G$，$\lvert S\rvert=r$） | 所选位点集合 |

# 4 问题一：

## 4.1 问题分析

本问要求根据分析需求，将位点的碱基对数据编码，达到最方便分析的效果。

本文首先分析了位点的碱基比例，碱基与疾病之间的关系，根据分析需求，发现此数据有两种适合的编码方式。第一种是由位点总采样数据分布决定的连续数值{0,1,2}编码方式，第二种是把原始数据的碱基对当做类别，采用独热编码(One-hot)的编码方式，将特征扩充三倍，每列均为{0,1}二进制编码。

## 4.2 基于位点碱基分布的编码方式

假设位点中，是以某碱基的个数影响患疾病的风险，例如位点 W 中，有三种碱基对组合，{C/C,C/T,T/T}，碱基 T 会提高患疾病 A 的风险，那么 C/C 组合是最安全的，C/T 的组合会稍微提高患病的风险，则 T/T 组合会显著增加患病的风险，那么，可以将{C/C,C/T,T/T}三种组合，编码为{0,1,2}。

同一位点仅存在两种碱基的组合，则设两碱基为{X,Y}

基于位点碱基分布的编码方式为

1. 计算碱基在位点出现的频次 $freq(X), freq(Y)$
2. 若 $freq(X) \geq freq(Y)$，则将 X/X 编码为 0，X/Y 编码为 1，Y/Y 编码为 2。
3. 若 $freq(X) < freq(Y)$，则将 X/X 编码为 2，X/Y 编码为 1，Y/Y 编码为 0。

## 4.3 独热编码(One-hot)编码方式

基于位点碱基分布的编码方式，是在更多的某碱基提高患病风险的假设上实现的，然而，通过本文对数据集的分析统计，发现有大量的位点，不符合这个原则，比如对于一个由 CC，CT，TT 组成的位点，CT 的患病概率可能比 CC，TT 都高。

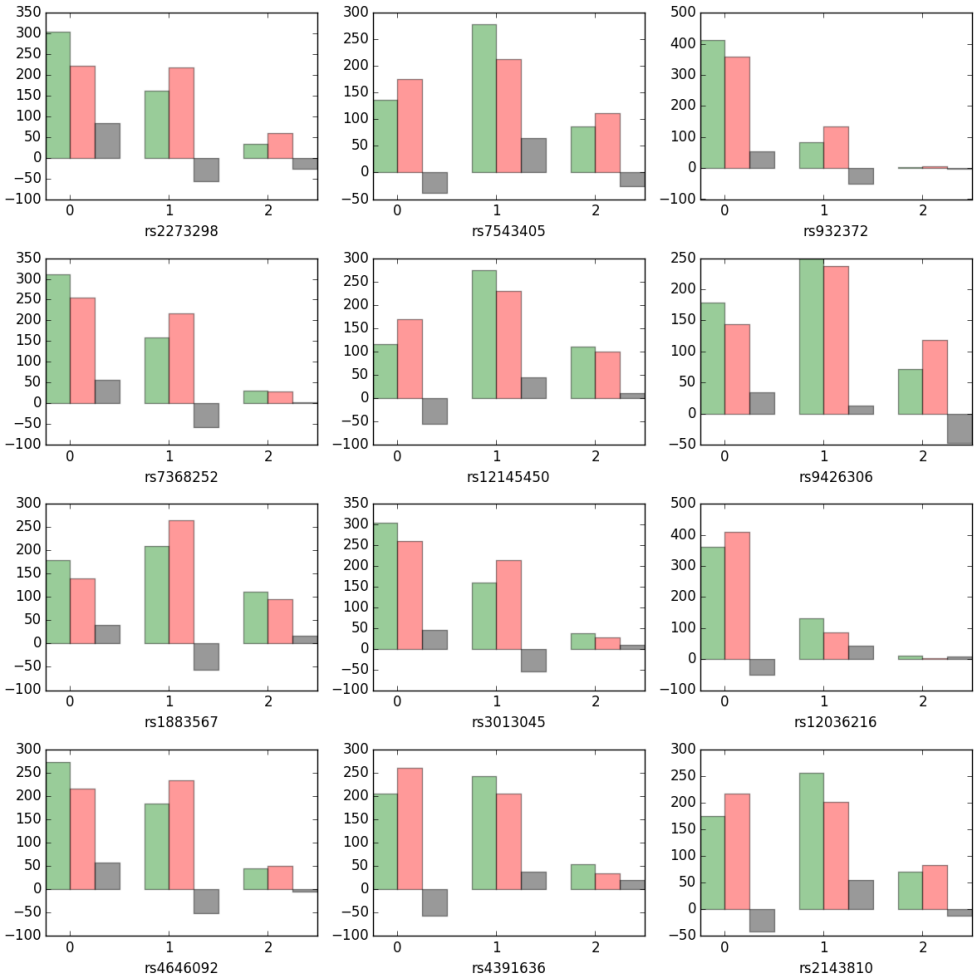


图 4-1 遗传位点碱基分布

由图 4-1 可见，rs7543405，rs12145450，rs214380 位点可以发现，这些位点中，由两种不同的碱基所组成的样本，患病概率最高。在这种场景下，本文采用了独热编码的编码方式。这种编码方式可以避免{0,1,2}连续数值编码所遇到的问题。

同一位点仅存在碱基对的三种组合，设为 $XX,XY,YY$

独热编码的编码过程如下：

1. 获取位点的三种编码组合，$XX,XY,YY$ 。
2. 将位点样本为 XX 的碱基对编码为[1,0,0]
3. 将位点样本为 XY 的碱基对编码为[0,1,0]
4. 将位点样本为 YY 的碱基对编码为[0,0,1]

独热编码的好处主要有：

1. 解决了分类器不好处理属性数据的问题
2. 一定程度上起到扩充特征的作用

# 5 问题二：

## 5.1 问题分析

本问题可以归结为单遗传疾病 A 与位点的相关性分析问题。核心在于通过对位点的非参数假设检验，过滤与疾病无关的位点，获取与疾病具有显著相关性的位点，然后从显著相关的位点中，选择出真正重要的位点，排除假阳性的位点。

文本的特征选择工作主要以下步骤组成：

1. 观察数据集，设计假设检验方案，通过假设检验过滤掉不显著相关的特征。
2. 引入噪声，通过模型特征选择的方法获得特征的重要性评分，通过迭代的方式过滤掉评分不如噪声的变量，最终得到高质量的相关位点。
3. 建立评估模型，为结果评估做准备。
4. 评估结果，将第 2 步得出的相关位点数据导入评估模型，通过预测准确率评估位点的相关性。

## 5.2 模型的建立与改进

### 5.2.1 回归随机森林

随机森林于 2001 年，由 Leo Breiman 提出，通过自助法重抽样技术，由随机向量$\theta$构成组合模型$\{h(X,\theta_k,k=1,...,p)\}$。预测变量为数值型变量，生成的随机森林，为多元非线性回归分析模型。随机森林通过求$k$棵树的$\{h(X,\theta_k)\}$的平均值，来进行预测，其中，形成的随机森林的训练集各自独立[1]。

给定训练集数据$\mathbb{L}=\{(\mathbf{X}_1,Y_1),...,(\mathbf{X}_N,Y_N)\}$，数字 N 是$\mathbb{L}$中的样本个数，一个回归

随机森林以如下方式创建：

步骤 1：原始训练集数据样本量为$\mathbb{L}$,通过 bootstrap 有放回地随机抽取 K 个自助样本集$\{\mathbb{L}_1,\mathbb{L}_2,...\mathbb{L}_K\}$，并由此构建 K 棵回归树，每次 bootstrap 抽样未被抽到的样本，组成了 K 个袋外数据集（out-of-bag samples），作为随机森林的测试样本。

步骤 2：从$\mathbb{L}_k$中生成回归树$T_k$。在每个节点，根据$\sum_{x_i\in t}(Y_i-\overline{Y}_t)/N(t)$决定分割，其中，$N(t)$是目标数量，$\overline{Y}_t$是在 t 节点处所有$Y_i$的平均值。

步骤 3：将生成的回归树组成随机森林回归模型进行预测，以$\hat{Y}^k$作为树$T_k$以输入$\mathbf{X}$的预测值，K 棵树的随机森林的回归预测结果由公式 1 计算

$$\hat{Y}=\frac{1}{K}\sum_{k=1}^{K}\hat{Y}^k \tag{1}$$

既然每棵树是从袋内样本生成的，那么它仅用了数据样本$\mathbb{L}$中三分之二的样本。另外三分之一的样本称作袋外样本，袋外样本(OOB)用于估计计算误差。

### 5.2.2 随机森林回归的变量重要性评价

对于给定的训练集数据$\mathbb{L}$，和一个回归随机森林模型 *RF*，使用基于permutation 随机置换的残差均方减少量进行衡量，其具体过程为：[2]

(1) 每一个自助样本建立一个回归树模型，其中$\mathbb{L}_k^{oob}$为第 k 棵树的袋外样本。对于给定的$\mathbf{X}_i \in \mathbb{L}_k^{oob}$，使用树$T_k$来预测$\hat{Y}_i^k$，函数为$\hat{f}_i^k(\mathbf{X}_i)$，得到 K 个袋外数据的残差均方$MSE_1, MSE_2, \cdots, MSE_K$。

(2) 变量$\mathbf{X}_i$在 K 个袋外样本中随机置换，形成新的袋外测试样本，然后用已建立的随机森林对新的袋外样本进行预测，与第一步的计算方法相同，得到随机置换后的 OOB 残差均方，重复 P 次得到矩阵

$$\begin{bmatrix} MSE_{11} & MSE_{12} & \cdots & MSE_{1K} \\ MSE_{21} & MSE_{22} & \cdots & MSE_{2K} \\ MSE_{31} & MSE_{32} & \cdots & MSE_{3K} \\ \vdots & \vdots & \vdots & \vdots \\ MSE_{P1} & MSE_{P2} & \cdots & MSE_{PK} \end{bmatrix}$$

(3) 用$MSE_1, MSE_2, \cdots, MSE_K$与矩阵对应的第 *i* 行向量相减，平均后再除以标准误差则得到变量$\mathbf{X}_i$的 OOB 评分，即公式 2：

$$score_i = \left(\sum_{j=1}^{K} (MSE_j - MSE_{ij}) / K\right) / SE, (1 \le i \le P) \tag{2}$$

(4) 比较有和没有 permutation 的均方残差(MSR)，以$\mathbf{X}_i$预测特征 *j*，其中，$MSR_i = \frac{1}{M_i} \sum_{k \in M_i} (\hat{f}_i^k(\mathbf{X}_i) - Y_i)^2$，$MSR_i^j = \frac{1}{P} \sum_{p=1}^{P} (\hat{f}_i^{p,j}(\mathbf{X}_i) - Y_i)^2$。

(5) 设$\Delta MSR_i^j = max(0, MSV_i^j - MSR_i)$，特征 j 的重要性评分为$IMP_j = \frac{1}{N} \sum_{i \in \mathbb{L}} \Delta MSR_i^j$。为了使重要性评分标准化，本文通过公式 3 获得原始重要性评分

$$VI_j = \frac{IMP_j}{\sum_l IMP_l} \tag{3}$$

通过公式给出的原始重要性评分，可以对变量的重要性做出排名。

### 5.2.3 基于假设检验的特征评估

本文需要在众多特征中，把重要的特征挑选出来。本文首先使用 Welch 假设检验的评分来和变量重要性评分进行比较，假设检验中最无关的噪音变量称为“阴影”特征。阴影特征对于本预测目标是不具有任何预测能力的。这样，任何

特征的，只要在随机森林的重要性评分中，小于阴影特征，那么也可以认为其不够重要。如果某特征的重要性评分高于阴影特征，那么可以认为其为重要特征[3]。

表格 5-1 M 个真实特征和 M 个阴影特征经过 R 次迭代组成的重要性评分矩阵

| 迭代 | $VI_{\mathbf{X}_1}$ | $VI_{\mathbf{X}_2}$ | … | $VI_{\mathbf{X}_M}$ | $VI_{\mathbf{A}_{M+1}}$ | $VI_{\mathbf{A}_{M+2}}$ | … | $VI_{\mathbf{A}_{2M}}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | $VI_{\chi_{1,1}}$ | $VI_{\chi_{1,2}}$ | … | $VI_{\chi_{1,M}}$ | $VI_{\alpha_{1,(M+1)}}$ | $VI_{\alpha_{1,(M+2)}}$ | … | $VI_{\alpha_{1,2M}}$ |
| 2 | $VI_{\chi_{2,1}}$ | $VI_{\chi_{2,2}}$ | … | $VI_{\chi_{2,M}}$ | $VI_{\alpha_{2,(M+1)}}$ | $VI_{\alpha_{2,(M+2)}}$ | … | $VI_{\alpha_{2,2M}}$ |
| ⋮ | ⋮ | | | | | | | ⋮ |
| R | $VI_{\chi_{R,1}}$ | $VI_{\chi_{R,2}}$ | … | $VI_{\chi_{R,M}}$ | $VI_{\alpha_{R,(M+1)}}$ | $VI_{\alpha_{R,(M+2)}}$ | … | $VI_{\alpha_{R,2M}}$ |

本文从拓展的数据集中，构建了随机森林模型 $RF$ 。根据章节 5.2.2 的 permutation 重要性评分，本文用 RF 为$2M$ 个特征计算$2M$ 各重要性评分，重复 R 遍来计算 R 遍重要性评分。表 5-1 表示 M 维输入数据特征和 M 个阴影特征，经过 permutation 方法得到的重要性评分矩阵。

从阴影特征的重要性评分中，本文提取出每行的最大值，然后作为对比样本 $V^* = max(A_{ri}),(r = 1,...R, i = M+1,...2M)$ 。对每个数据特征 $\mathbf{X}_i$ ,通过公式 4 计算 $t$ 统计量

$$t_i = \frac{\bar{X}_i - \bar{V}^*}{\sqrt{(s_1^2 + s_2^2)/R}} \tag{4}$$

其中， $s_1^2$ 和 $s_2^2$ 是两样本变量的无偏估计量。

对于显著性测试，公式 4 中， $t_i$ 的分布自由度 $df$ 由公式 5 计算得出

$$df = \frac{(s_1^2/n1 + s_2^2/n_2)^2}{(s_1^2/n1)^2/(n_1-1) + (s_2^2/n_2)/(n_2-1)} \tag{5}$$

此处 $n_1 = n_2 = R$ 。

### 5.2.4 高维数据保证质量的的特征选择模型

随机森林的重要性评分只给出了一个位点重要性排名，然而，由于全基因组分析数据的天然噪声特性，很难从单一的排名中选取出最重要的位点。为了更好地在树的每个节点进行子空间选择，本文首先要把高信息量的位点与噪音位点中区分开，然后，高信息量的位点根据统计量，被划分到两个组中，当对位点子空间抽样时，包含高信息量位点的子空间就会被保证更好地划分树的节点。

在第一步，本文先建立 R 个随机森林来获得重要性评分，然后本文使用 Fisher 假设检验方法，或者 M 个 p 值小于阈值 $\theta$ （默认设置为 0.05）的位点，同时，将 M 个 p 最值高的（无相关性）的噪声位点作为“阴影”位点，然后将 2M 的数据通过作为随机森林模型选择，获得所有的重要性评分进入比较环节，因为阴影位点不具有预测能力，如果一个位点位点是真正重要的，那么它在多次排列组

合的预测中的，重要性总会高于阴影位点，由此本文可以得到真正重要的位点SNP[4][5]。

本模型基本思想基于以下事实：

1. 高维数据的特征空间中含有许多冗余特征，甚至噪声特征，这些特征一方面可能会降低分类或者回归的精度，另一方面会大大增加学习及训练的时间及空间复杂度。[6]
2. 随机森林适合于分析变量数大于样本数的数据，而且可以对每个变量的重要性进行评分。
3. 随进森林利用自助法抽样，考虑袋外数据的预测，不易产生过拟合问题，在高维数据的降维方面具有独特的优势。

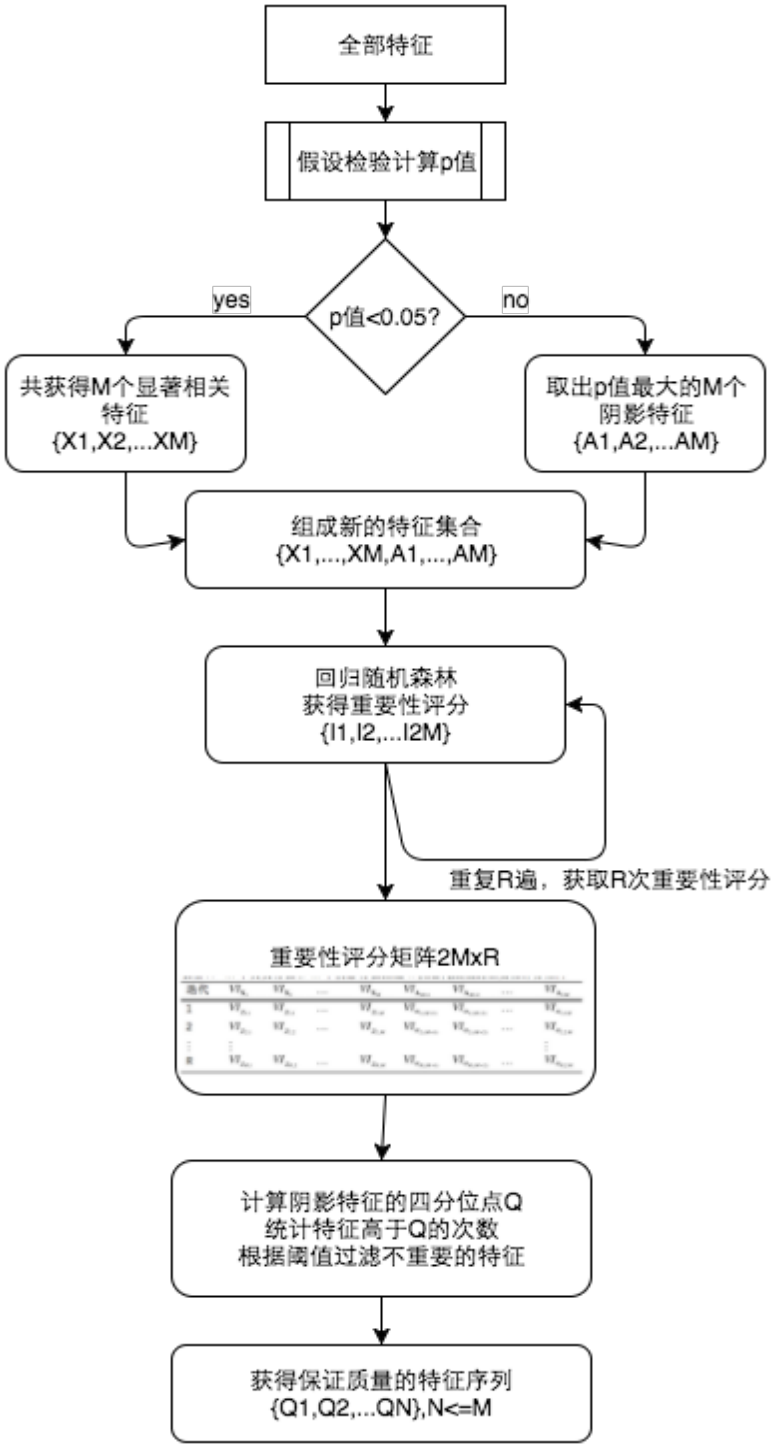


图 5-1 保证质量的特征选择算法流程

质量系数的定义：在 R 次迭代中，变量 $\mathbf{X}_i$ 重要性评分高于阴影变量统计量四分位点 Q 的次数为 S，则变量地方 $\mathbf{X}_i$ 的质量系数为：

$$score_Q = S / R$$

保证质量的的特征选择算法如下：

（1） 通过给定的 $\mathbb{L}$，生成额外的数据 $\mathbb{L}^e$，通过假设检验特征评估的方法，用阴影特征，将数据扩充到 2M 维度。

（2） 从$\mathbb{L}^e$数据中，建立随机森林模型，为预测特征和阴影特征和计算 R 遍原始重要性评分，把每次计算的重要性评分提取，作为对比组。
（3） 对每个预测性特征，按照公式 4 来计算 t 统计量。
（4） 按照公式 5 计算$df$的自由度。
（5） 给定 t 统计量和$df$，计算所有预测性特征的 p 值。
（6） 给定阈值$\lambda$，将重要的特征和不重要的特征分类两组，$\mathbf{X}_{low}$和$\mathbf{X}_{high}$。
（7） 有放回地抽样训练集$\mathbb{L}$，生成 K 份 Bagged 样本$\mathbb{L}_1,\mathbb{L}_2,...,\mathbb{L}_K$。
（8） 对于每个样本$\mathbb{L}_k$，通过以下方法生成回归树$T_k$

a) 在每个节点，随机选择$m=\lfloor\sqrt{M}\rfloor$个特征，然后区别为$X_{low}$和$X_{high}$，然后用子空间特征作为划分节点的候选人。
b) 每个数以不确定的方式生成，在每个叶节点，保持叶节点所有$Y$值。
c) 为每棵树和森林的 out-of-bag 样本，计算每个$X_i$的权重。

（9） 给定一个概率$\tau$,$\alpha_l$和$\alpha_h$，使得$\alpha_h-\alpha_l=\tau$，计算相应的四分位点$Q_{\alpha l}$和$Q_{\alpha h}$。
（10）统计每个特征$\mathbf{X}_i$重要性评分小于 Q 的次数，生成频率列表，取频率高于阈值$\lambda$的特征为最重要的特征。

## 5.3 模型结果与分析

模型取 R=1000，质量系数阈值$\lambda$=0.8，输入数据，得到重要特征序列如下：

表格 5-2 重要特征表

| 序号 | 位点名称 | P-value | 变量重要性评分 | 质量系数 |
|---|---|---|---|---|
| 0 | rs2273298 | 0.000054 | 4.1204 | 1.000 |
| 1 | rs7543405 | 0.000212 | 2.3546 | 1.000 |
| 2 | rs932372 | 0.000241 | 2.2025 | 1.000 |
| 3 | rs7368252 | 0.00069 | 2.0523 | 1.000 |
| 4 | rs12145450 | 0.000545 | 1.9901 | 0.997 |
| 5 | rs9426306 | 0.00043 | 1.9396 | 0.993 |
| 6 | rs1883567 | 0.001597 | 1.9025 | 0.997 |
| 7 | rs3013045 | 0.001762 | 1.8918 | 0.995 |
| 8 | rs12036216 | 0.000422 | 1.8087 | 0.991 |
| 9 | rs4646092 | 0.001559 | 1.7825 | 0.995 |
| 10 | rs4391636 | 0.000761 | 1.7574 | 0.981 |
| 11 | rs2143810 | 0.00219 | 1.7413 | 0.986 |
| 12 | rs880801 | 0.001595 | 1.737 | 0.994 |
| 13 | rs1541318 | 0.002574 | 1.7301 | 0.994 |
| 14 | rs2250358 | 7.123748 | 1.6809 | 0.968 |
| 15 | rs15045 | 0.001619 | 1.5682 | 0.971 |
| 16 | rs9659647 | 0.004429 | 1.5424 | 0.970 |
| 17 | rs1138333 | 0.005114 | 1.5276 | 0.977 |
| 18 | rs2807345 | 0.001361 | 1.4591 | 0.936 |
| 19 | rs2095518 | 0.006949 | 1.4562 | 0.936 |

| | | | | |
|---|---|---|---|---|
| 20 | rs5746051 | 0.001216 | 1.4508 | 0.933 |
| 21 | rs1891419 | 0.007308 | 1.4469 | 0.954 |
| 22 | rs11121557 | 0.006651 | 1.4105 | 0.921 |
| 23 | rs11573253 | 0.004286 | 1.4018 | 0.906 |
| 24 | rs12036552 | 0.006814 | 1.3944 | 0.925 |
| 25 | rs6699113 | 0.007949 | 1.3801 | 0.934 |
| 26 | rs7555715 | 0.007989 | 1.3059 | 0.866 |
| 27 | rs7543486 | 0.01127 | 1.3038 | 0.884 |
| 28 | rs11247865 | 0.002671 | 1.3017 | 0.856 |
| 29 | rs10779765 | 0.008226 | 1.3012 | 0.899 |
| 30 | rs1009113 | 0.011461 | 1.2904 | 0.883 |
| 31 | rs3765380 | 0.011446 | 1.2901 | 0.857 |
| 32 | rs2473246 | 0.008666 | 1.2767 | 0.860 |
| 33 | rs10916825 | 0.018797 | 1.2682 | 0.874 |
| 34 | rs707472 | 0.004337 | 1.2681 | 0.836 |
| 35 | rs2038095 | 0.009391 | 1.2653 | 0.839 |
| 36 | rs1148455 | 0.007822 | 1.2577 | 0.833 |
| 37 | rs946758 | 0.009226 | 1.2511 | 0.827 |
| 38 | rs12044299 | 0.025281 | 1.2402 | 0.831 |
| 39 | rs2788891 | 0.006578 | 1.2337 | 0.817 |
| 40 | rs7534822 | 0.012947 | 1.2279 | 0.818 |

由表 5-2，可以发现其中 rs2273298、rs7543405、rs932372、rs7368252 四个位点在 1000 次迭代中，重要性永远大于阴影特征，由此可以说，四个特征的质量是具有保证的。第 5-40 个特征，重要性在绝大部分的迭代中，也大于阴影特征。

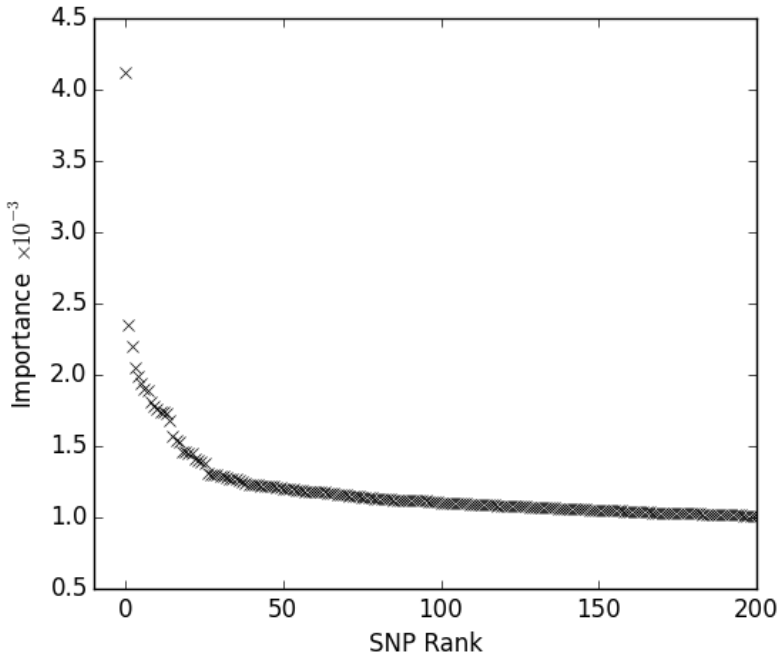


图 5-2 遗传位点重要性排名曲线

图 5-2 展示的是遗传位点重要性从大到小排名的曲线，可以看出，小部分遗传位点具有高的变量重要性，大部分遗传位点具有很低的变量重要性，可以解释为长尾效应。从图中可以看出，大约在排名 40 的位点，变量重要性骤减。

## 5.4 评估模型及评价指标

为了评估特征选择模型的有效性，本文采用了支持向量机模型(SVM)，基于树的模型，线性模型、和贝叶斯模型。

### 5.4.1 支持向量机模型

给定训练向量 $x_i \in \mathbb{R}^p, i=1,...,n$，和向量 $y \in \{1,-1\}^n$，支持向量机解决了以下最优化问题

$$\begin{aligned} &\min_{w,b,\ell} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}\zeta_i \\ &\text{subject to } y_i(\omega^T\phi(x_i)+b) \geq 1-\zeta_i \\ &\zeta_i \geq 0, i=1,...,n \end{aligned} \tag{6}$$

它的对偶是

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ &\text{subject to } y^T\alpha = 0 \\ &0 \leq \alpha_i \leq C, i=1,...,n \end{aligned} \tag{7}$$

其中，$e$ 是全为 1 的向量，$C>0$ 是上限，$Q$ 是 $n$ 乘 $n$ 的半正定矩阵，$Q_{ij} = y_i y_j K(x_i,x_j)$，其中 $K(x_i,x_j) = \phi(x_i)^T\phi(x_j)$ 是核，将训练向量，通过 $\phi$ 方程隐性地将其映射到了一个更高维度的空间。
决策方程为

$$sgn(\sum_{i=1}^{n} y_i\alpha_i K(x_i,x) + \rho) \tag{8}$$

### 5.4.2 决策树模型

决策树构建的机构步骤如下：

(1) 开始，所有记录看做一个节点
(2) 遍历每个变量的每一种分割方式，找到最好的分割点
(3) 分割两个节点 N1 和 N2
(4) 对 N1 和 N2 分别执行第 2-3 步，直到每个节点足够“纯”为止

其中，量化纯度由公式 9 计算：

$$Entropy = -\sum P(i) * \log_2^{P(i)} \tag{9}$$

除了决策树，本文还将其组合模型随机森林作为评估的模型之一。

### 5.4.3 朴素贝叶斯模型

本文评估模型采用了伯努利朴素贝叶斯模型。

伯努利朴素贝叶斯模型实现了朴素贝叶斯的训练和分类的算法，用于以多变量伯努利分布的数据。可能有多个特征，但是每个特征都是假设为二元的伯努利变量。

伯努利朴素贝叶斯的决策规则基于公式 10

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i) \tag{10}$$

### 5.4.4 评价指标

本题本质上是二分类问题，根据测试集真实类别，与模型预测类别组合划分为真正例、假正例、真反例、假反例四种情形，令 TP、FP、TN、FN 分别表示对应的样本数，分类矩阵的混淆矩阵如表 5-3 所示：

表格 5-3 混淆矩阵

| 真实情况 | 预测结果 | |
|---|---|---|
| | 正例 | 反例 |
| 正例 | TP(真正例) | FN(假反例) |
| 反例 | FP(假正例) | TN(真反例) |

本文采用以下三个指标评估结果：
准确率 P、召回率 R 和 F1 度量，由公式 11 计算：

$$P=\frac{TP}{TP+FP},$$
$$R=\frac{TP}{TP+FN},\tag{11}$$
$$F1=\frac{2\times P\times R}{P+R}$$

## 5.5 结果验证

对于模型一产生的结果，本文将产生的 40 个特征作为变量，对样本的健康状况进行预测。本文的测试用，采用了支持向量机模型、基于树的模型、线性模型、贝叶斯模型、和组合模型。

因为原始数据集的数据量偏小，本文采用 10 折交叉验证的方法，各种模型的测试集准确率、召回率、F1 度量如表 5-4 所示。

表格 5-4 评估模型验证指标表

| 模型 | | 准确度 | 召回率 | F1 度量 |
|---|---|---|---|---|
| SVM | SVM(rbf) | 0.656 | 0.6425 | 0.6492 |
| | SVM(linear) | 0.668 | 0.6506 | 0.6592 |
| Tree-based | DecisionTree | 0.552 | 0.5692 | 0.5605 |
| | ExtraTrees | 0.704 | 0.6937 | 0.6988 |
| | RandomForest | 0.712 | 0.7208 | 0.7164 |
| Linear | SGDClassifier(L1) | 0.711 | 0.6985 | 0.7046 |
| | SGDClassifier(L2) | 0.714 | 0.6728 | 0.6928 |
| | LogisticRegression(L1) | 0.711 | 0.7129 | 0.7119 |
| | LogisticRegression(L2) | 0.706 | 0.7034 | 0.7047 |
| Bayes | BernoulliNB | 0.721 | 0.7055 | 0.7137 |
| Ensemble | AdaBoost | 0.722 | 0.7013 | 0.7114 |
| | GradientBoosting | 0.692 | 0.6945 | 0.6932 |

由表 5-4 可以看出，随机森林模型对于预测疾病 A 的表现最优秀，F1 度量达到了 71.64%，组合模型 AdaBoost 的准确率最高，达到了 72.2%。在 40 位点征的预测中，可以通过预测的准确率证明这 40 个位点为与遗传疾病 A 相关的致命位点。

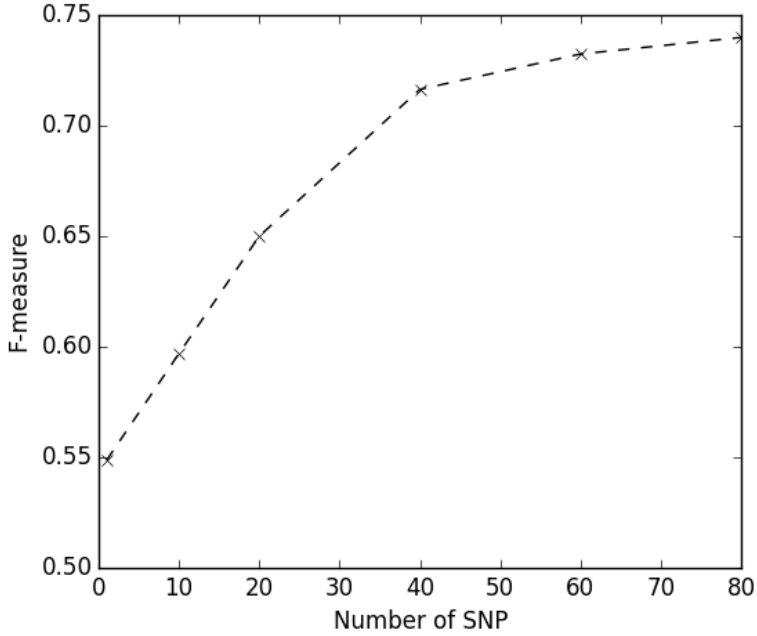

图 5-3 遗传位点数量与预测 F 度量关系

随着质量系数阈值的降低，可以获得更多的位点，从而达到更高的预测准确度，图 5-3 表示 F 度量与参与预测的位点数量的关系，可以得知，从 1 个到 40 位个点的过程中，F 度量快速上升，在 40 个输入特征时，达到 72%。从 40 个到 80 个位点的过程中，F 度量缓慢上升最终在 80 个位点时，达到了 74%。可以确认，保证质量的特征选择模型的确选出了真正重要的致病位点。

# 6 问题三：

## 6.1 问题分析

近年来，随着全基因组关联研究的深入，国内外医学遗传学研究已广泛关注到多基因复杂性状疾病机制的深入挖掘，试图通过 GWAS 数据中蕴含的基因数据中获取丰富信息，并探索潜在的疾病机制。本问题在问题二的基础上继续探索基因和疾病的关联性，并要求找出与前文相同的遗传疾病 A 的最有可能相关的一个或几个基因。

在第二问中要求挑选出与遗传疾病 A 最为相关的位点。针对单个位点的分析往往是从统计学角度出发，从众多基因中挑选出最具有统计学意义或假阳性最小的位点，然而可能存在如下的两个问题，其一是挑选出的位点即使具有很强的统计学意义，然而由于样本数量的限制，这种统计学的强相关性也可能是由于存在某种巧合所致；其二，那些统计学意义相对较弱的位点虽然没有通过多重检验，但也不意味着就完全没有关联，其仍然有可能是假阴性，导致了那些实际易感的基因被遗漏了。

与第二问单独考虑各个位点不同，基因水平的关联分析需要考虑同一个基因上 位点的数量，一般可以推测同一个基因中与某个疾病相关联的位点的数量越多，占基因所有位点的比例越高，该基因可能与这种疾病更加相关；同时，由于位点之间的连锁不平衡结构，即同一个基因上的不同的 A、B 两个位点一般不是独立关联的（或称它们具有非随机关联性），因此也并非具有越多关联的 SNP 的基因就与疾病的关联性越大。通常衡量这种连锁平衡程度的方法如下：假设 A、B 两个位点编码为 a、b，在群体中出现的概率分别为$P(a)$、$P(b)$，则可以用$D=P(ab)-P(a)\times P(b)$来衡量这种不平衡程度。位点之间的连锁不平衡结构等因素的存在，导致在遗传学的概念上具有一定的复杂性，也给统计方法带来了许多挑战性。

目前，以基因为单位进行的疾病关联分析，主要方法包括生物网络分析和生物通路分析,基于此的分析必须在分析前将基因上全部或者部分位点的遗传关联结果综合起来（也即基因水平的关联分析）。无论是生物网络分析方法还是生物通路分析方法，都需要将基因中位点作为一个集合一起进行考虑，实际中可以将基因理解成为该基因中所包含的若干个位点的集合，而遗传疾病与基因的关联性可以由基因中所包含的位点的全集或其子集合最终表现出来。

针对本问题，本章将基因作为与疾病关联性分析的研究对象，将一个基因上的多个位点看做自变量$X$,将疾病的表现看出因变量$Y$,通过在$X$和$Y$之间建立起回归方程，并以决定系数$R^2$（coefficient of determination）作为统计量，判断基因与疾病的关联结果。通常，可以采用的回归分析方法包括线性回归（Linear regression）、岭回归(Ridge regression)、逻辑回归(Logistic regression)和典型相关分析等。采用回归方法的优势在于能够直接对疾病在基因的水平上（或多个位点）进行关联分析，而不需要建立在位点水平上的关联分析，同时，该方法计算复杂度较低，能够在较短时间内对大量基因的关联性进行估计，同时，以决定系数作为评判的依据具有较好的理论基础，能够直接说明自变量（基因）对因变量（疾病）的解释程度。

## 6.2 模型建立

由于位点并非独立的，疾病的发生也不仅仅是某一个位点的单独作用，因此需要在基因水平上展开分析，挖掘出与疾病最为相关的基因。本章针对位点之间的相关性以及基因和疾病之间的相关性分别建立了两个模型。针对位点之间的相关性，本章采用互信息模型衡量同一个基因的不同位点或不同基因的位点之间是否存在连锁不平衡效应展开分析，说明了相同基因的位点之间连锁不平衡现象的存在。因此并不能直接对每个基因的位点和疾病之间的相关性的显著程度直接进行简单的统计加和，而需要考虑基因内不同位点之间的交互关系。基于此，针对基因和疾病之间的相关性建立了回归分析模型。

### 6.2.1 衡量位点之间相关性的互信息模型

首先介绍衡量位点之间相关性的互信息模型如下。互信息（Mutual Information）是信息论里一种有用的信息度量，它可以看成是一个随机变量中包含关于另一个随机变量的信息量，因此可以用来衡量两个随机变量之间的关联程度，即给定一个随机变量后，另一个随机变量不确定性的削弱程度。如果两个变量之间存在确知的关系，即知道其中一个变量的值能够有很高的概率正确推出另外一个变量的值，则两个变量之间存在很强关联性，即它们之间的互信息很高，反则两个变量如果是完全独立的，已知一个变量的取值对推测另外一个毫无用处，则它们之间的互信息为零[7]。

假设两个离散随机变量 $X$ 和 $Y$ 的概率分布分别为 $p(x)$ 、$p(y)$ 。它们的联合概率分布为 $p(x,y)$ ,则他们之间的互信息定义为：

$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{12}$$

由此可知，当变量 $X$ 和 $Y$ 完全独立时，即 $p(x,y)=p(x)p(y)$ ，有

$$\log\left(\frac{p(x,y)}{p(x)p(y)}\right)=\log 1=0 \tag{13}$$

因此有 $I(X;Y)=0$ 。

在衡量位点之间相关性时，变量 $X$ 和 $Y$ 分别代表了两个位点各自的三种组合方式。两个位点之间的互信息越高，它们之间的相关性越大，在后一小节，本文将通过实验的方式验证相同基因的位点之间的互信息可能明显大于不同基因的位点之间的互信息，因此不能忽略他们之间的相关性(连锁不平衡效应)。

### 6.2.2 衡量基因与疾病相关性的回归模型

#### 6.2.2.1 基因与疾病相关性分析常用方法

首先对调研得到的主流基因水平上的关联分析做一个简单的综述，而后将详细地介绍本文所建立的回归分析模型。

目前常见的基因水平上的关联分析方法主要包括最显著 SNP 法（Best SNP method）、组合法（Combination method）、回归分析方法（Regression analysis）、主成分分析法（Principal component analysis）、傅里叶分析法（Fourier analysis）等[8]。

✧ 最显著 SNP 法或最大 OR 法

所谓最显著 SNP 法是指把某个基因中最显著的 SNP 的显著水平（P-value 最小）作为该基因的显著水平，即 $P=\min\{P_i\},i=1,2,...,n$。最显著 SNP 法又称 Best SNP 或 Most-significant SNP 方法。其优点是思想朴素、算法简单。OR(odd ratio)又名机会比，优势比，交叉乘积比（Cross-product Ratio），当基因上多个 SNP 对应的病人的对照样本量一致时，越小的 P-value 对应越大的 OR，因此最大 OR 法和最显著 SNP 法等价，只有当基因分型缺失时，样本量的不一致可能导致 P-value 和 OR 存在差异。

✧ 次显著 SNP 法

不同于最显著 SNP 法，次显著 SNP 法通过次小而非最小的 P-value 作为评判该基因关联性的判据。这样做的好处在于有效避免了 GWAS 中随机关联的 SNP（即只是由于样本的偶然因素所引发的统计相关性，而并非实际上存在关联），采用次显著 SNP 方法相比最显著 SNP 方法更加稳健保守，能够有效降低 GWAS 中所存在的假阳性结果。

✧ 组合法

组合法主要包括 Fisher 组合法和截断乘积法。在组合法中，一个基因上的多个位点均假设不存在连锁平衡关系，也即所有位点的 P-value 都是相互独立切均匀分布的，因此可以用过组合的方法计算一个基因内所有位点的遗传关联结果。Fisher 组合法将一个基因上所有的位点的 P-value 直接进行相乘，得到检验统计量 $X=-2\ln(\prod_{i=1}^{n}p_i)=\sum_{i=1}^{n}-2\ln(p_i)$ 服从自由度为 2n 的 $\chi^2$ 分布。截断乘积法（TPM）是建立在 Fisher 组合法之上，区别于 Fisher 组合法，截断乘积法只对 P-value 小于阈值 $\tau$ 的 SNP 的 P-value 进行连乘。

✧ 回归分析方法

将一个基因上的多个位点看做自变量 $X$,将疾病的表现看出因变量 $Y$,通过在 $X$ 和 $Y$ 之间建立起回归方程，并以决定系数 $R^2$，方程检验的 P-value 或似然值等作为疾病和基因关联判据的统计量。回归分析方法包括线性回归、岭回归、逻辑回归和典型相关分析等一些列方法。这些方法均建立在线性回归的数学模型至上，但由于具有不同的假设，因此对应了不同的表达形式，具有了不同的应用场景，并适用于不同类型的数据。本章主要采用了线性回归、岭回归、逻辑回归的方法。具体模型的建立将在后文详细阐述。

✧ 主成分分析和傅里叶分析方法

由于回归方法中普遍存在多重共线性的问题（虽然岭回归方法能够针对多重共线性问题进行处理，但仍然存在基因水平关联分析中的高自由度问题，即基因中存在较多的位点使得检验效能较低方程不稳定）。主成分分析方法采用了降维的思路，将一组可能相关的变量重新组合成少量的相互独立的变量，因此能够良好的解决多重共线性问题。而由于基因和表型直接可能存在非线性关系，因此后续也发展出核主成分分析，即采用核函数对原始变量空间进行变化，将线性空间转换成非线性空间。主成分分析通常用于处理高维数据（高维空间中普遍存在众多相关联的变量）。同傅里叶分析同主成分分析一样也能处理高维数据，主要基于傅里叶变化，将多个位点的基因分型结果转化为傅里叶变换的主成分,并对这些主成分进行加权评分得到检验统计量。傅里叶分析具有可靠性高和检验效能高的特点。[9]

目前，基因水平的关联分析还是不断探索阶段，还不是特别成熟，存在不少亟待解决的问题。下面将重点介绍本文所采用的回归分析方法。

#### 6.2.2.2 线性回归

为了解释逻辑回归模型和岭回归模型，首先简要阐述线性回归模型。回归分析中最基本的方法是线性回归。统计学中，线性回归采用最小二乘函数对多个自变量和因变量之间的关系进行建模，即将因变量表示成为多个自变量之间的线性组合，线性组合之前的系数称为回归系数。只有一个自变量的回归称为简单回归，多于一个自变量的回归称为多元回归。

线性回归的理论模型如下。给一个随机样本$(Y_i, X_{i1},...,X_{ip}), i=1,...,n$，线性回归模型除了假设回归因子$Y_i$和回归量$X_{i1},...,X_{ip}$之间存在线性关系，还采用增加一个误差项$\varepsilon_i$来反应除去回归量$X_{i1},...,X_{ip}$外其他因素的影响。因此一个多元线性回归模型可以表示成为下面的形式：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_m X_{im} + \varepsilon_i, i = 1,...,n \tag{14}$$

为了简化，可以采用矩阵简化为：

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{15}$$

对$\beta$的估计方法之一是选择是的残差平方和达到最小：

$$\hat{\beta} = \arg\min_{\beta} ||\mathbf{Y} - \mathbf{X}\beta||^2 = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{k=1}^{m} \beta_k x_{ik})^2 \tag{16}$$

当$\mathbf{X}$为满秩时有：

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{17}$$

#### 6.2.2.3 方差分析

为了评估回归拟合的好坏程度，并解释因变量和自变量之间的相关性，需要进行方差分析。在方差分析（ANOVA）中，可以对因变量的方差进行分解，并归属到不同的自变量来源，在回归分析中，方差分析具有以下概念。

- 总平方和 SST(sum of squares for total)

$$\text{SST}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^{\,2}-\frac{1}{n}(\sum_{i=1}^{n}y_i)^2,\quad \text{其中}\ \bar{y}=\frac{1}{n}\sum_i y_i \tag{18}$$

- 回归平方和 SSReg(sum of squares for regression)

$$\text{SSReg}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{19}$$

- 残差平方和 SSE(sum of squares for error)

$$\text{SSE}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{20}$$

由以上有：

$$\text{SST}=\text{SSReg}+\text{SSE} \tag{21}$$

- 决定系数$\mathrm{R}^2$

$$\mathrm{R}^2=\frac{\text{SSReg}}{\text{SST}}=1-\frac{\text{SSE}}{\text{SST}} \tag{22}$$

决定系数$\mathrm{R}^2$用来解释因变量的方差有多少能够由自变量来解释，其大小决定了相关的密切程度。$\mathrm{R}^2$越接近 1，表示相关性越高，反之越接近 0，相关性越底。

#### 6.2.2.4 逻辑回归

逻辑回归属于对数线性模型，线性回归模型进行回归学习，而逻辑回归模型本质上是一种分类模型。

由于患病与否是二分类任务，自变量$\mathbf{X}$为向量，取值为实数，因变量$y$取值为 0（无病）或 1（有病），线性模型产生的预测值$z=\beta\mathbf{X}+b$是实数，需要转换为 0/1 值，最理想的是“单位阶跃函数”，即若预测值$z$大于零判为正例，小于零判为反例，临界值则任意判断。但单位阶跃函数不连续，不具有可微性，因此可以用如下的对数几率函数替代，图示如下。

$$y=\frac{1}{1+e^{-z}} \tag{23}$$

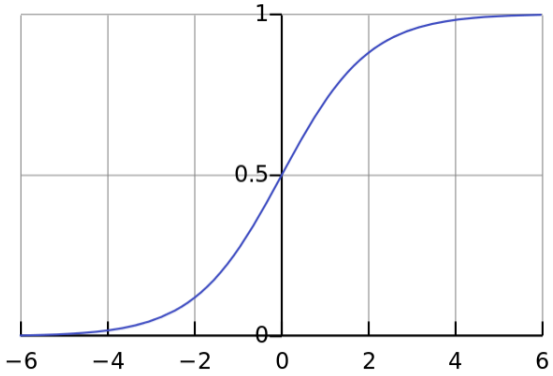

图 6-1 对数几率函数

将 $z=\beta\mathbf{X}+b$ 带入上式得：

$$y=\frac{1}{1+\exp(\beta\mathbf{X}+b)} \tag{24}$$

可变化为：

$$\ln\frac{y}{1-y}=\beta\mathbf{X}+b \tag{25}$$

其中 $\ln\frac{y}{1-y}$ 反映了 $\mathbf{X}$ 作为正例的相对可能性，因此称为“对数几率”。

针对二项逻辑回归模型有如下的条件概率分布：

$$\begin{aligned} P(y=0|\mathbf{X})&=\frac{1}{1+\exp(\beta\mathbf{X}+b)}\\ P(y=1|\mathbf{X})&=\frac{\exp(\beta\mathbf{X}+b)}{1+\exp(\beta\mathbf{X}+b)} \end{aligned} \tag{26}$$

对于给定数据集 $\{(\mathbf{X}_i,y_i)\}_{i=1}^m$，回归模型最大化“对数似然”：

$$\ell(\beta,b)=\sum_{i=1}^m\ln p(y_i|\mathbf{X}_i;\beta,b)=\sum_{i=1}^m(-y_i\beta^T\hat{\mathbf{X}}_i+\ln(1+e^{\beta^T\hat{\mathbf{X}}_i})) \tag{27}$$

上式是关于 $\beta$ 的高阶可导连续凸函数，根据凸优化理论，可以采用数值优化算法如梯度下降法、牛顿法等求取其最优解。

$$\beta^*=\arg\min_{\beta}\ell(\beta) \tag{28}$$

估计基因和疾病的相关性，可以针对基因中的位点的全集或其子集和疾病做逻辑回归，对模型检验的 P-value 可以作为基因关联结果，而决定系数 $R^2$ 可以用来说明该基因与疾病关联性的强弱。

#### 6.2.2.5 岭回归

这里的岭回归是指采用岭回归的方法实现的岭回归分类器，本质上还是一种分类模型，类似于逻辑回归。

由于在做基因和疾病的相关性分析时，每个基因上通常都包含有较多个数的位点，这些位点通常别变化为 0、1 和 2 或采用独热编码成 0、1 等，由于自变量之间存在相关性（连锁不平衡特性），在回归分析中容易产生多重共线性问题，进而导致采用最大似然估计的回归系数不稳定，影响了回归方程的建立。不同于逻辑回归，为了解决多重共线性问题，岭回归通过获取回归系数的最小有偏估计，来确保回归方程的稳定。

因此，采用岭回归的分析方法，不仅能够获取基因和疾病的关联性，更能较好的估计每个位点的相关性强弱，具有比逻辑回归更高的检验效能。

在进行特征选择时，一般有三种方式：子集选择、收缩方式（又称为正则化）、维数缩减。岭回归属于收缩方式，即在平方误差的基础上增加了正则项：

$$\sum_{i=1}^{n}(y_i-\beta\mathbf{X}_i)^2+\lambda\beta\beta^T \tag{29}$$

在逻辑回归和岭回归（分类）中经常使用的是伪决定系数，例如 McFadden 决定系数。

### 6.3 模型结果与分析

基于问题二位点和疾病的相关性分析，以及筛选出的可能与疾病 A 相关联的 400 多个位点（通过 P-value 检测），首先统计出了基因中包含可疑位点的数量的分布统计，分布如图 6-2 所示，横坐标是基因中包含位点的数量，纵坐标是对应的基因数量，由此可见，大多数基因只包含少量位点，但仍然有相当比重的基因包含两个或以上的位点。

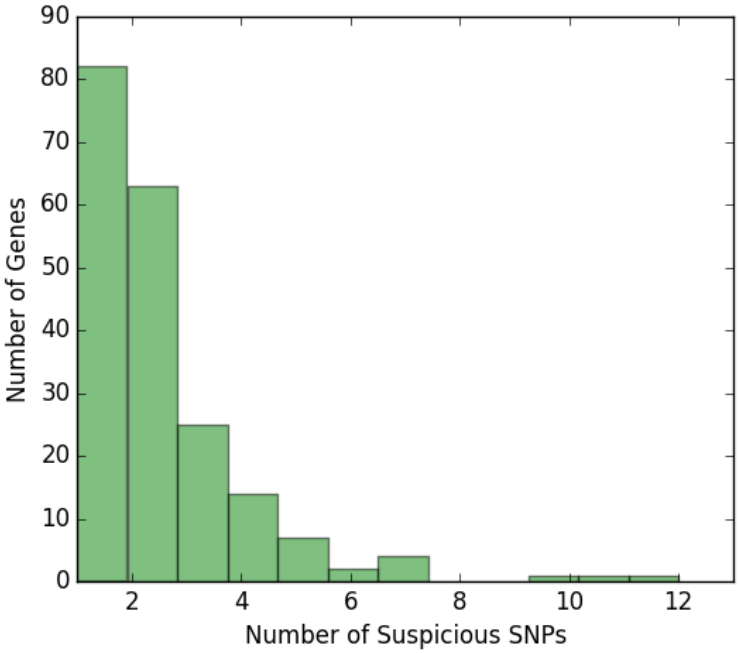

图 6-2 基因内位点数量分布图

采用互信息模型，衡量了相同基因之间以及不同基因之间的位点的相关性，结果如图 6-3 所示。其中横、纵坐标均是疑似位点按照原顺序（染色体上的顺序）进行排序，其中颜色较深部分的互信息较大，对应的位点之间可能存在某种关系，由结果的分布可知，对角线附近的位点存在互信息较大部分，而非对角线部分的互信息均较小，而角线附近的位点距离较近，经过验证除去极少数基因，互信息较大的基因均存在与相同的染色体上，验证了基因中位点存在连锁不平衡性质，因此不能假设基因中位点相互独立。

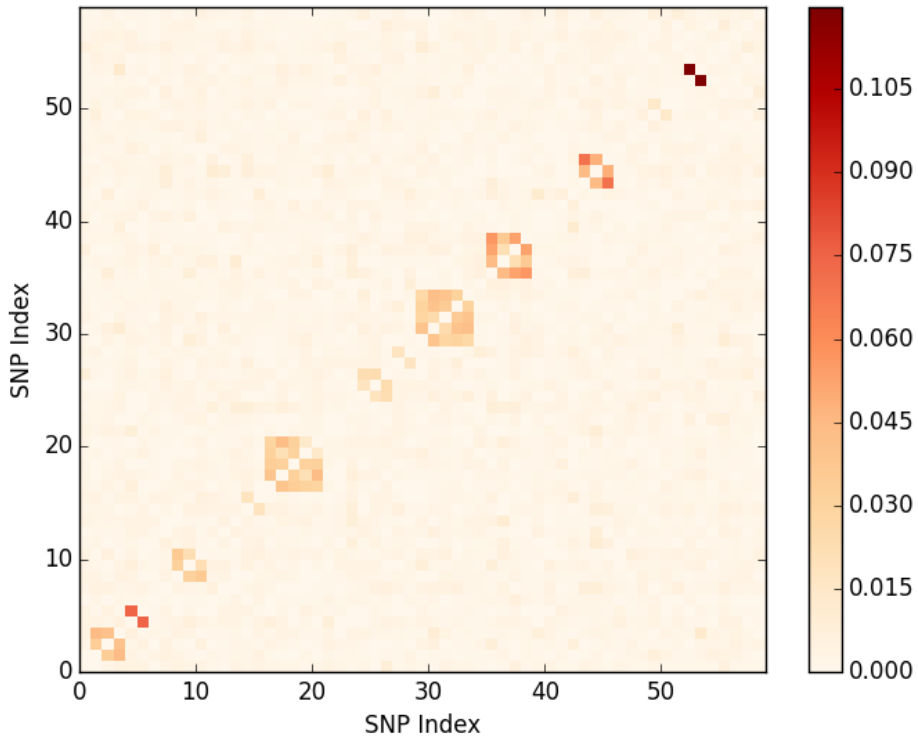


图 6-3 可疑位点互信息矩阵

如表 6-1 所示是采用衡量基因与疾病相关性的两种不同回归模型得到的最优可能与 A 疾病相关的前 20 的基因（按照决定系数 $R^2$ 由大到小进行排序），其中只有一个不同，剩下的 19 个仅排序存在一些差异，如表 6-1 所示，所获取的基因具有两个特点，其一是所包含的可疑位点数量一般较多，其二是其中包含的可疑位点的重要性评分的最高分一般较高。

表格 6-1 逻辑回归与岭回归结果对比

| | 逻辑回归 | | | | 岭回归 | | | |
|---|---|---|---|---|---|---|---|---|
| 序号 | 基因名称 | 决定系数 | 位点数量 | 位点最高评分 | 基因名称 | 决定系数 | 位点数量 | 位点最高评分 |
| 1 | gene_293 | 0.196 | 12 | 1.306 | gene_162 | 0.198 | 7 | 1.741 |
| 2 | gene_162 | 0.194 | 7 | 1.741 | gene_55 | 0.196 | 11 | 2.052 |
| 3 | gene_265 | 0.19 | 10 | 2.355 | gene_265 | 0.194 | 10 | 2.355 |
| 4 | gene_55 | 0.186 | 11 | 2.052 | gene_114 | 0.178 | 5 | 1.163 |
| 5 | gene_114 | 0.178 | 5 | 1.163 | gene_254 | 0.176 | 4 | 1.302 |
| 6 | gene_254 | 0.176 | 4 | 1.302 | gene_293 | 0.174 | 12 | 1.306 |
| 7 | gene_30 | 0.166 | 5 | 0.999 | gene_78 | 0.17 | 6 | 1.218 |
| 8 | gene_78 | 0.166 | 6 | 1.218 | gene_30 | 0.166 | 5 | 0.999 |
| 9 | gene_102 | 0.166 | 2 | 4.120 | gene_102 | 0.166 | 2 | 4.120 |
| 10 | gene_217 | 0.162 | 7 | 1.459 | gene_217 | 0.162 | 7 | 1.459 |
| 11 | gene_128 | 0.16 | 2 | 1.265 | gene_283 | 0.158 | 5 | 1.180 |
| 12 | gene_131 | 0.158 | 4 | 1.020 | gene_131 | 0.158 | 4 | 1.020 |
| 13 | gene_168 | 0.154 | 4 | 1.228 | gene_22 | 0.154 | 5 | 1.041 |
| 14 | gene_113 | 0.154 | 6 | 1.109 | gene_168 | 0.152 | 4 | 1.228 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 15 | gene_283 | 0.152 | 5 | 1.180 | gene_113 | 0.15 | 6 | 1.109 |
| 16 | gene_22 | 0.152 | 5 | 1.041 | gene_163 | 0.144 | 4 | 1.164 |
| 17 | gene_35 | 0.15 | 4 | 1.757 | gene_35 | 0.144 | 4 | 1.757 |
| 18 | gene_163 | 0.148 | 4 | 1.164 | gene_260 | 0.14 | 4 | 1.143 |

如图 6-4 所示是对基因按照决定系数$R^2$由大到小进行排序后的决定系数$R^2$的分布，由此可见曲线存在两个拐点，其中左边第一个拐点靠左的基因决定系数较大，而往右决定系数变化趋缓，本文认为左边的 20-40 个是更有可能的基因集合。

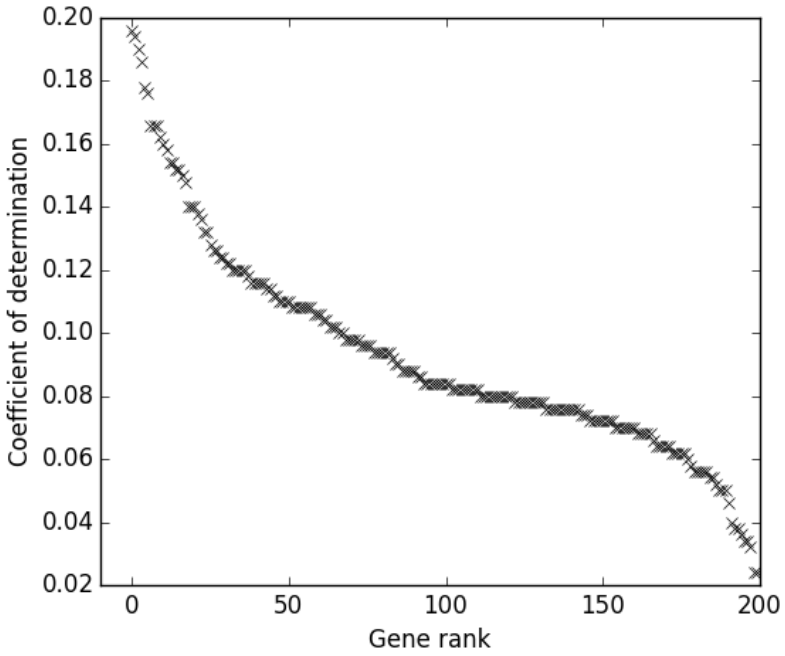


图 6-4 基因的决定系数排名曲线

如图 6-5 所示是对所选中的 20 个决定系数$R^2$最大的基因的特性的分布，包含两个维度，即基因中包含可疑位点的数量和包含的可疑位点的重要性评分的最高分，说明包含可疑位点数量较多，或其中包含的可疑位点的重要性评分的最高分较高的基因更有可能和疾病相关，同时只包含一个可疑位点的基因均为被选中，一定概率避免了由于某种巧合因素所导致的假阳性，而同时也存在包含多个可疑位点而未被选中的情况，这可能是因为这些位点具有较高的互信息，回归系数不够显著。

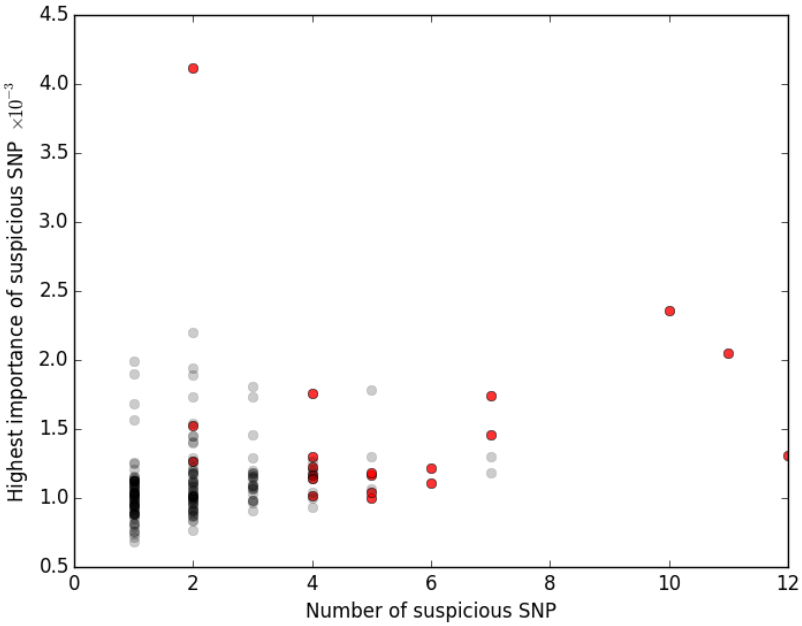


图 6-5 基因特性分布图

## 6.4 结果验证

最后通过与问题二相同的验证方法，验证了本文获取的最可能的 20 个基因为最有可能和疾病 A 相关的基因。采用 10 折交叉验证的方法，多种模型的测试集准确率、召回率、F1 统计量如表 6-2 所示。采用这些基因获取了趋近于问题二评估的准确率。

表格 6-2 评估模型验证指标表

| 模型 | | 准确度 | 召回率 | F1 统计量 |
|---|---|---|---|---|
| SVM | SVM(rbf) | 0.636 | 0.6668 | 0.651 |
| | SVM(linear) | 0.626 | 0.676 | 0.65 |
| Tree-based | DecisionTree | 0.62 | 0.5949 | 0.6072 |
| | ExtraTrees | 0.691 | 0.6873 | 0.6891 |
| | RandomForest | 0.702 | 0.7108 | 0.7064 |
| Linear | SGDClassifier(L1) | 0.679 | 0.6916 | 0.6852 |
| | SGDClassifier(L2) | 0.649 | 0.6827 | 0.6654 |
| | LogisticRegression(L1) | 0.691 | 0.7092 | 0.6999 |
| | LogisticRegression(L2) | 0.654 | 0.6834 | 0.6683 |
| Bayes | BernoulliNB | 0.689 | 0.6951 | 0.692 |
| Ensemble | AdaBoost | 0.684 | 0.684 | 0.684 |
| | GradientBoosting | 0.656 | 0.681 | 0.6682 |

# 7 问题四：

## 7.1 问题分析

上文中研究了单位点与单性状的关系，多位点（基因）与单性状的关系。在实际中，个体往往表现出多种性状而非单一性状，这些性状与多个位点相关。此外，各个性状之间并非完全独立，而是会有一定的相关性。因此把相关的性状（或疾病）看成一个整体往往更有意义，探寻与他们相关的位点。即比起找出与单性状最相关的位点，本文更希望找出同时与多个性状相关的位点。由于这些位点同时与多个性状相关，因此着重研究这些位点，对研究人类性状表达（或遗传病）有着更本质的意义。

本文用 $G=\{\alpha_i | i=1,\ldots,m\}$ 表示题目中所包含的所有位点的集合；用 $L=\{\lambda_i | i=1,\ldots,n\}$ 表示题目中所包含的所有性状的集合。本文希望找出与这些性状同时最相关的 $r$ 个位点 $S\subseteq G$。值得强调的是，本文希望本文找出的位点集合 $S$，与所有的性状整体最相关，而不是找出位点集合 $S'$，仅仅与其中的某个性状相关性较高，而与其他性状的相关性较低。

本文把该问题作为基于**多标签分类问题（Multi-label Classification）**的特征选择（Feature Selection）来建模。多标签分类问题是当今机器学习领域的一个研究热点，已经在视频语义标注[10][11][12]、音乐情感分类[13][14][15]、新闻类别标记[17][18][19]等问题中得到了广泛的应用。近年来，多标签分类也越来越多地出现在基因功能组的相关研究中[20]。接下来，本文首先介绍多标签分类问题的相关方法，以及它在基因位点识别问题中的应用。之后，结合本文数据的特点建立模型，包括分析各个性状的相关性（彼此独立还是相关），并结合多标签分类特征选择思想建模。最后本文给出实验结果，包括选出的位点名称及相关特征、对比实验结果，以及深入的分析和论证。

## 7.2 相关工作

多标签分类问题（Multi-label Classification）以机器学习算法为基础，属于机器学习中的有监督学习问题（Supervised Learning）。首先通过对已有数据的学习建立模型，再通过训练好的模型对新的实例（Instance）进行分类。与单标签分类问题不同，多标签分类问题的标签是一个集合，对每个实例而言，不是单一的类别，而是一组类别向量。多标签分类的任务就是要为每个实例标注出与之相关的所有标签，而达到多标签管理的目的。多标签分类的一般流程如下：

- 对训练集(Training Set)学习，建立分类模型；
- 利用学习好的分类模型，对于一个待预测样本，在给定的**标签集合**范围内输出该样本所属的所有标签；

已有的多标签分类方法主要包含两大类：**问题转换法(Problem Translation, PT)**和**算法适应法(Algorithm Adaptation)**。问题转换法的主要思想是将多标签分类问题，转换为一个或多个单标签分类学习的问题。由于核心思想是转换，因此

该算法并不受限于各种单标签学习的算法。本文可以使用很多已有的单标签学习算法，沿用这些传统的方法解决多标签分类的问题，例如支持向量机(Support Vector Machine)、朴素贝叶斯分类器(Naïve Bayesian )、K 近邻方法(K Nearest Neighbor)。算法适应法则是直接在现有的单标签分类学习方法，使之能够处理多标签分类的问题。常见的算法适应法包括决策树(Decision Tree)、Boosting 算法、概率图模型(Probabilistic Graphical Model)等。表 7-1 比较了几种常见的多标签分类方法的优点和缺点。

表格 7-1 主要多标签数据挖掘方法比较

| 类型 | 算法 | 模型或方法 | 特点 | 不足 |
|---|---|---|---|---|
| PT 问题转换 | ECC | BR | 考虑了标签之间的相互依赖关系 | 需要构建多条分类器链削弱链内顺序对准确率造成的影响 |
| | RPC | RPC | 数据集不易产生偏斜,采用投票的方式产生标签 | 类别种类很多时,构造的子分类器过多,需要经验数据确定筛选标签的阈值 |
| | CRL | RPC | 加入人工校准标签区分相关标签和不相关标签 | 类别标签很多时,构造的子分类器过多 |
| | LP | LP | 充分考虑了标签之间的依赖关系,简单有效地将多标签问题转化为单标签问题 | 新产生的标签数量较多,同时有可能造成个别标签样本过少导致偏斜,无法准确预测训练集中从未出现过的标签组合 |
| AA 算法适应 | C4.5 | C4.5 | 可以从数据集中学习到一些精确而有意义的多标签分类规则 | 不能解决完全的分类问题 |
| | BoostTexter | Boosting | 通过迭代对样本权重进行修正,全局优化损失函数,直至收敛到最优情况 | 迭代复杂度比较高,可能出现过拟合 |
| | SVM | SVM、BR | 采用两轮迭代训练,考虑了标签间的相互关联信息,基于剪枝混淆矩阵消除相似标签干扰 | 本质上使用 BR 转换,也存在数据偏斜问题 |
| | CRF | PM | 充分考虑了类别独立性和多标签稀疏性,将类别依赖关系直接运用到算法中 | 特征属性必须采用统一的权重和类型,使用的二分剪枝法缺乏灵活性 |

## 7.3 模型建立

### 7.3.1 问题定义

给定：

- $G=\{\alpha_i | i=1,\ldots,m\}$ 表示所包含的所有位点的集合；

- $L=\{\lambda_i \mid i=1,\ldots,n\}$ 表示所包含的所有性状的集合；
- $D=\{\mathbf{X}_i,\mathbf{Y}_i \mid i=1,...,m\}$ 表示多标签数据的集合，其中$\mathbf{X}_i$是第$i$个样本的特征向量，$Y_i(Y_i \subseteq L)$是第$i$个样本所属的标签集合；

找出：

- 位点集合$S$（$S \subseteq G$，$|S|=r$)与这些性状同时最相关；

### 7.3.2 多标签分类建模

(1) 多标签特征分类方法选择

多标签分类方法主要包含两大类：**问题转换法(Problem Translation, PT)**和**算法适应法(Algorithm Adaptation)**。其中问题转换法主要思想是将多标签分类问题，转换为一个或多个单标签分类学习的问题，较为适合个标签之间相互较为独立，相关性较小的数据集；而算法是应法直接在现有的单标签分类学习方法，使之能够处理多标签分类的问题，较为适合各个标签之间相关性较大的数据。

本文首先对数据样本中的 10 种性状$\lambda_i(i=1,\ldots,n)$进行相关性分析，使用皮尔逊相关系数(Pearson Correlation Coefficient Coefficient)刻画性状间两两的相关性，如公式 30 所示。

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left(\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right)}{\sigma_X\sigma_Y} \tag{30}$$

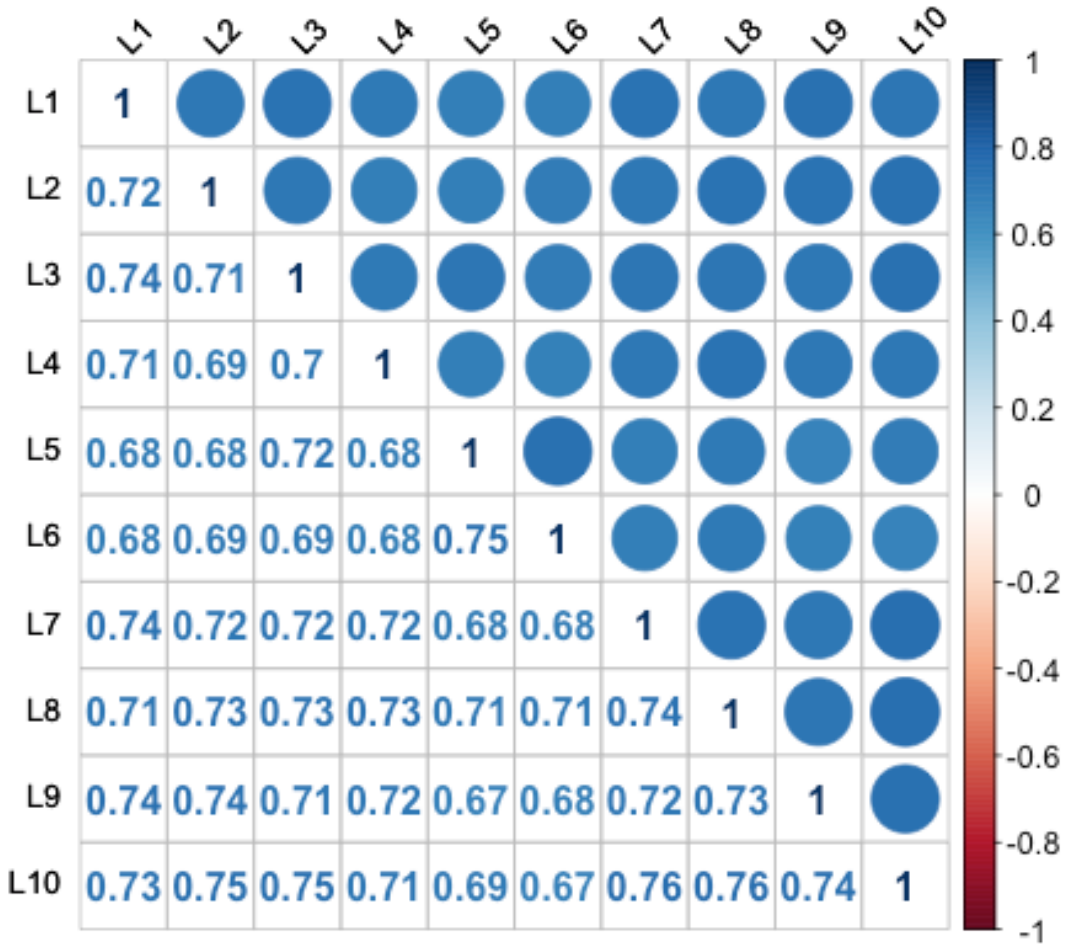


图 7-1 多性状相关性分析

图 7-1 所示展示出了性状之间的相关性。蓝色表示正相关，红色表示负相关，数值越大表示相关性越强。从图 7-1 可见，10 个性状两两相关性多在$r=0.7$左右(大于$r_0=0.5$)，各个性状彼此之间强相关。因此，本文采用算法适应法的方法**(Algorithm Adaptation)**。

(2) 多标签随机森林模型(Multi-label Random Forest, MLRF)

结合样本数据的位点特征，以及性状之间的相关性，本文以随机森林(Random Forest)算法为基础进行建模。主要考虑以下两点：

- 每个性状的碱基编码为类别型数据，随机森林(决策树)最擅长处理类别型特征，而支持向量机等只能处理数值型特征，会造成一部分信息丢失；
- 样本数据中样本数(1000)远小于位点数(9445)，相比于决策树，处理高维数据随机森林具有较大优势，不会产生过拟合;

随机森林是一种集成算法(ensembling)，每个弱分类器是一棵决策树(Decision Tree)。它流程图树结构，其中每个内部节点表示在一个属性上的测试,每个分枝代表一个测试输出，而每个树叶节点代表类或类分布。C4.5 是一种常见的决策树构造算法，它具有如下优点：用信息增益来选择属性；在树的构造过程中进行剪枝；能够完成对连续属性的离散化处理；能够对不完整数据进行处理。

信息熵(Information Entropy)是度量样本集合纯度最常用的一种指标，如公式 31 所示，其中$D$表示样本集合，$p_k$表示$D$中第 k 类样本所占的比例，$Ent\left(D\right)$的值越小，则$D$的纯度越高。

$$Ent\left(D\right)=-\sum_{k=1}^{|y|}p_k\log_2 p_k \tag{31}$$

信息增益(Information Gain)表示信息熵的一种变化，如公式 32 所示。一般而言，信息增益越大，则意味着使用属性 a 来进行划分所获得的“纯度提升”越大。

$$Gain\left(D,a\right)=Ent\left(D\right)-\sum_{v=1}^{V}\frac{\left|D^v\right|}{\left|D\right|}Ent\left(D^v\right) \tag{32}$$

针对多标签分类问题，Clare 和 King[12]改进了该算法得到了 Multi-label-C4.5(ML-C4.5)，算法中允许叶节点为一个标签的集合并修改了相应的信息熵计算公式，如 33 所示。其中$p\left(\lambda_j\right)$是属于标签$\lambda_j$的概率，$q\left(\lambda_j\right)=1-p\left(\lambda_j\right)$。

$$Entropy\left(D\right)=-\sum_{j=1}^{q}\left(p\left(\lambda_j\right)\log p\left(\lambda_j\right)+q\left(\lambda_j\right)\log p\left(\lambda_j\right)\right) \tag{33}$$

随机森林是 Bagging 的一个扩展变体。RF 是一组决策树的集成，其中每一棵决策树都是由随机抽取的样本和特征训练而成。每棵决策树按下面方法生成，首先把叶节点分为左右两个子节点，之后根据分割条件把该节点中的数据点分给两个子节点。分割条件通常是由数据点的特征与阈值比较决定。而所选择的特征以及阈值则通过优化代价函数决定，例如基尼指数。整个训练过程知道所有的决策树训练完成或者基尼系数不再下降。针对单标签学习的随机森林方法，本文把每棵决策树的构建方法从 C4.5 改进为 ML-C4.5，记得到了多标签随机森林分类模型，Random Forest of ML-C4.5 (RFML-C4.5)。

### 7.3.3 算法设计

建立好模型之后，本文结合 MLRF 模型设计位点选择算法。首先根据 p-value 进行假设检验，筛掉与各个性状无关的位点，得到位点集合$S^{'}(S^{'}\subseteq G)$。之后根据 MLRF 计算出$S_1$中各个位点的 Importance 值，筛选出位点集合$S(S\subseteq S^{'})$。之后对筛选出的位点集合进行验证。

通过假设检验筛选位点集合，由于每个样本有 10 个性状，筛选方法与单性状有所不同。本文对每个性状，分别计算每个位点的 p-value，对每个性状筛选出相应的位点集合$S_i(i=1,\ldots,10)$，$S^{'}=\bigcup_{i=1}^{10}S_i$。本文统计了$S^{'}$之中每个性状被 10 个性状选中的次数（最少的被选中 1 次，最多的被选中 10 次），如图 7-2 所示。总共有大约 1200 个性状至少被一个性状选中，绝大多数的性状被选中的次数较少(不多于 3 次)，少数性状被选中的次数最多(大于 3 次)，本文挑选出被 3 个以上性状选中的特征作为$S^{'}$(约 200 个性状)。

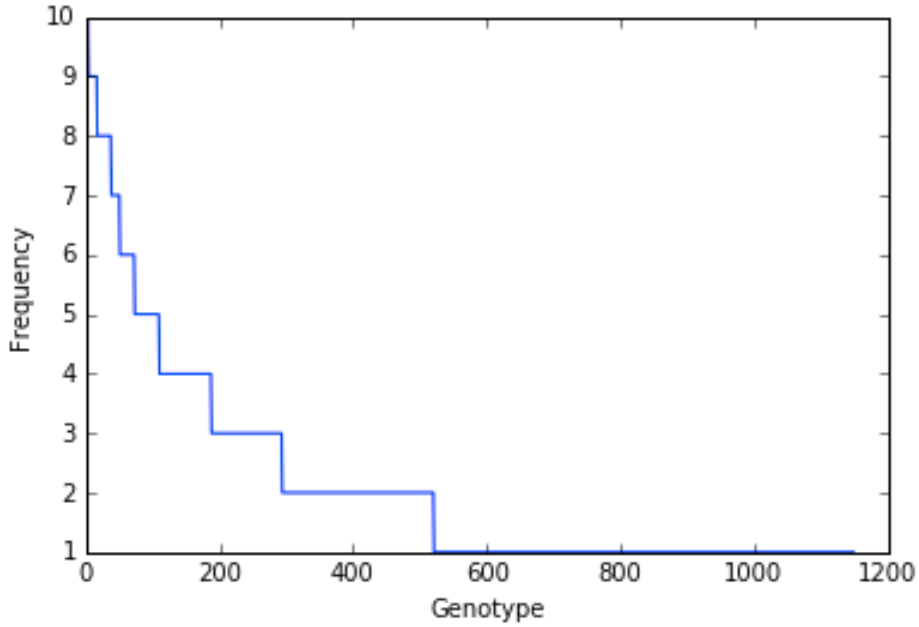


图 7-2 多性状显著基因频次

得到$S'$之后，本文按照多标签分类模型 MLRF，设计算法得到$S$：

（1） 给定阈值$\lambda$，将重要的特征和不重要的特征分类两组，$\mathbf{X}_{low}$和$\mathbf{X}_{high}$。

（2） 有放回地抽样训练集$\mathbb{L}$，生成 K 份 Bagged 样本$\mathbb{L}_1,\mathbb{L}_2,...,\mathbb{L}_K$。

（3） 对于每个样本$\mathbb{L}_k$，通过以下方法生成回归树$T_k$

a) 在每个节点，随机选择$m=\lfloor\sqrt{M}\rfloor$个特征，然后区别为$X_{low}$和$X_{high}$，然后用子空间特征作为划分节点的候选人。

b) 每个数以不确定的方式生成，在每个叶节点，保持叶节点所有$Y$值。

c) 为每棵树和森林的 out-of-bag 样本，计算每个$X_i$的权重。

（4） 给定一个概率$\tau$,$\alpha_l$和$\alpha_h$，使得$\alpha_h-\alpha_l=\tau$，计算相应的四分位点$Q_{\alpha l}$和$Q_{\alpha h}$。

（5） 统计每个特征$\mathbf{X}_i$重要性评分小于 Q 的次数，生成频率列表，取频率高于阈值$\lambda$的特征为最重要的特征。

## 7.4 模型结果与分析

根据以上算法，本文选出了与 10 个性状最相关的 40 个位点，位点的名称及其指标(被性状选中的频次、重要性)如表 7-2 所示。根据以上结果，可以发现其中 rs728340、rs2501275 位点被 10 个性状共同选中，并且重要性也在所有性状中排名靠前。由此可以说，性状的质量是具有保证的。

表格 7-2 相关位点及其指标

| 序号 | 位点名称 | 选中频次 | 重要性 |
|---|---|---|---|
| 1 | rs728340 | 10 | 0.009779251 |
| 2 | rs2501275 | 10 | 0.008948622 |
| 3 | rs35107626 | 9 | 0.009501513 |
| 4 | rs387232 | 9 | 0.008792639 |

| | | | |
|---|---|---|---|
| 5 | rs11247925 | 8 | 0.009308301 |
| 6 | rs12722898 | 9 | 0.009389097 |
| 7 | rs2247525 | 9 | 0.008754144 |
| 8 | rs2745252 | 9 | 0.008357437 |
| 9 | rs3218121 | 10 | 0.008963346 |
| 10 | rs7515988 | 8 | 0.008950375 |
| 11 | rs1000313 | 9 | 0.008960647 |
| 12 | rs780983 | 8 | 0.008561165 |
| 13 | rs1149046 | 7 | 0.008910562 |
| 14 | rs4073710 | 10 | 0.008575973 |
| 15 | rs4073395 | 8 | 0.007921746 |
| 16 | rs716325 | 9 | 0.008593907 |
| 17 | rs17257107 | 7 | 0.00856041 |
| 18 | rs11121821 | 8 | 0.008944244 |
| 19 | rs351617 | 9 | 0.008299916 |
| 20 | rs198372 | 9 | 0.008141652 |
| 21 | rs7538876 | 9 | 0.008550619 |
| 22 | rs6424069 | 9 | 0.008221771 |
| 23 | rs10737913 | 8 | 0.009035645 |
| 24 | rs12137141 | 8 | 0.008414144 |
| 25 | rs7549888 | 8 | 0.008520935 |
| 26 | rs560514 | 8 | 0.008492831 |
| 27 | rs11573253 | 8 | 0.008474957 |
| 28 | rs3767216 | 8 | 0.008507779 |
| 29 | rs16823542 | 8 | 0.008353859 |
| 30 | rs2501430 | 8 | 0.008238405 |
| 31 | rs4908602 | 8 | 0.008097221 |
| 32 | rs4654361 | 9 | 0.00855746 |
| 33 | rs7522419 | 8 | 0.008707694 |
| 34 | rs267700 | 8 | 0.008772823 |
| 35 | rs1739822 | 8 | 0.008201609 |
| 36 | rs4360511 | 8 | 0.008377488 |
| 37 | rs12065908 | 8 | 0.008874017 |
| 38 | rs2376723 | 8 | 0.008438132 |
| 39 | rs12758112 | 8 | 0.008872592 |
| 40 | rs8019 | 9 | 0.008724901 |

## 7.5 结果验证

对于选出来的 40 个位点，本文通过 MLRF 模型进行 10 折交叉验证，以准确率(accuracy)作为评估指标。如表 7-3 所示，最后一行，表示使用多标签模型的预

测结果，可见多标签模型对 10 个性状的预测准确率都比较高，均为 70%以上，10 个性状准确率的平均值达 72%。

为了说明本文模型的性能，本文进行了对比试验，与单标签模型进行了对比。对比试验中，针对 10 个性状中的每一种性状，本文采用问题二中的解决方案，筛选出 40 个特征，之后使用这 40 个特征，采用随机森林模型，对每一个性状 10 折建模预测，得到相应的准确率。例如，本文使用性状 1 进行单模型性状选择，得到了 40 个性状，然后使用这 40 个性状对 10 个性状分别做 10 折预测。观察表 7-3 第一行可见，通过性状 1 选出的 40 个性状，对自身预测准确率高达 75.9%，但对其他 9 个性状预测准确率较低，10 个性状准确率平均值较低(66.8%)。其他性状的单标签预测也类似。可见，通过多标签分类模型选出的位点集合 $S$，与所有的性状整体相关度最高；而通过单标签分类模型选出位点集合 $S'$，仅仅与其中的某个性状相关性较高，而与其他性状的相关性较低。

表格 7-3 多标签模型与单标签模型性状选择预测准确率评估

| 建模\结果 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 平均 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 性状 1 | **0.759** | 0.639 | 0.65 | 0.672 | 0.643 | 0.65 | 0.666 | 0.677 | 0.66 | 0.67 | 0.668 |
| 性状 2 | 0.642 | **0.747** | 0.638 | 0.624 | 0.635 | 0.649 | 0.633 | 0.649 | 0.672 | 0.671 | 0.656 |
| 性状 3 | 0.645 | 0.659 | **0.743** | 0.646 | 0.644 | 0.651 | 0.673 | 0.649 | 0.66 | 0.671 | 0.664 |
| 性状 4 | 0.665 | 0.66 | 0.655 | **0.77** | 0.65 | 0.666 | 0.67 | 0.688 | 0.666 | 0.677 | 0.6767 |
| 性状 5 | 0.654 | 0.634 | 0.64 | 0.646 | **0.742** | 0.664 | 0.642 | 0.634 | 0.653 | 0.637 | 0.655 |
| 性状 6 | 0.656 | 0.654 | 0.654 | 0.655 | 0.678 | **0.749** | 0.653 | 0.648 | 0.662 | 0.66 | 0.667 |
| 性状 7 | 0.67 | 0.654 | 0.669 | 0.685 | 0.643 | 0.652 | **0.76** | 0.686 | 0.688 | 0.689 | 0.68 |
| 性状 8 | 0.686 | 0.656 | 0.673 | 0.68 | 0.654 | 0.649 | 0.686 | **0.749** | 0.671 | 0.662 | 0.677 |
| 性状 9 | 0.667 | 0.677 | 0.674 | 0.66 | 0.651 | 0.631 | 0.676 | 0.67 | **0.757** | 0.666 | 0.673 |
| 性状 10 | 0.66 | 0.666 | 0.664 | 0.656 | 0.649 | 0.643 | 0.66 | 0.67 | 0.66 | **0.753** | 0.668 |
| **多标签** | 0.706 | 0.721 | 0.722 | 0.743 | 0.726 | 0.707 | 0.719 | 0.713 | 0.737 | 0.72 | **0.721** |

# 8 模型评价

对于问题 2，本文通过融合过滤式特征选择和包裹式特征选择，可以从高维数据的特征空间中，鉴别出真正重要的特征，本质上是通过引入噪声的方式，对不够重要的特征进行了过滤，非常适用于较高维度数据中的特征选择。

对于问题 3，衡量位点之间相关性的互信息模型具有计算简单、效率高等优点，具有良好的理论基础，并能够有效检测出位点之间连锁不平衡现象的存在。衡量基因与疾病相关性所采用的逻辑回归模型和岭回归模型考虑到了位点之间的相关性，采用决定系数 能够较好的衡量基因与疾病的相关性强弱，岭回归分类器相对于逻辑回归模型还考虑到了变量之间的多重共线性，更能较好的估计每个位点的相关性强弱，具有更高的检验效能，缺点在于要以损失部分信息、降低精度为代价。

对于问题 4，本文通过假设检验与多性状投票相结合的方式进行特征筛选，选择算法适应法(Algorithm Adaptation)，以 Random Forest of ML-C4.5 (RFML-C4.5)建模。该模型适合处理碱基编码这种类别型数据，适合处理强相关性状，而且处理高维数据不易过拟合，准确率较高。不足之处是当弱分类器数目过多时训练时间较长。

# 9 参考文献


[1] Qi, Yanjun. "Random forest for bioinformatics." *Ensemble machine learning*. Springer US, 2012. 307-323.

[2] 杨凯, 侯艳, and 李康. "随机森林变量重要性评分及其研究进展." (2015).

[3] Tung, Nguyen Thanh, et al. "Extensions to quantile regression forests for very high-dimensional data." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer International Publishing, 2014.

[4] Winham, Stacey J., et al. "SNP interaction detection with Random Forests in high-dimensional genetic data." *BMC bioinformatics* 13.1 (2012): 1.

[5] Nguyen, Thanh-Tung, et al. "Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests." *BMC genomics* 16.Suppl 2 (2015): S5.

[6] Botta V, Louppe G, Geurts P, et al. Exploiting SNP correlations within random forest for genome-wide association studies[J]. PloS one, 2014, 9(4): e93379.

[7] 陈峰, et al. "全基因组关联研究中的统计分析方法." *中华流行病学杂志* 32.4 (2011): 400-404.

[8] 罗旭红, 刘志芳, and 董长征. "基因水平的关联分析方法." *遗传* 35.9 (2013): 1065-1071.

[9] Gauderman, W. James, et al. "Testing association between disease and multiple SNPs in a candidate gene." *Genetic epidemiology* 31.5 (2007): 383-395.

[10]Mathew R B,Luo Jie-bo,Shen Xi-peng,et al.Learning multi-la-bel scene clasification[J].Patern Recognition,2004(37):1757-1771

[11]Zhang Min-ling,Zhou Zhi-hua.Ml-kNN:A lazy learning ap-proach to multi-label learning[J].Patern Recognition,2007 (40):2038-2048

[12]Xu Xin-shun,Jiang Yuan,Peng Liang,et al.Ensemble approach based on conditional random field for multi-label image and video annotation[C]// Procedings of the 19th ACM international conference on Multimedia.Scotsdale,Arizona,USA,2011: 1377-1380

[13] Li, Tao, and Mitsunori Ogihara. "Detecting emotion in music." *ISMIR*. Vol. 3. 2003.

[14] Sanden, Chris, and John Z. Zhang. "Enhancing multi-label music genre classification through ensemble techniques." *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011.

[15] 马宗杰. *Sparse 方法在多标签分类中的应用*. MS thesis. 浙江师范大学, 2014.

[16]Elisef A,Weston J.A kernel method for multi-labeled clasifi cation[J].Advances in Neural Information Procesing Systems, 2001(14):681-687

[17] Blockeel, Hendrik, et al. "Decision trees for hierarchical multilabel classification: A case study in functional genomics." *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer Berlin Heidelberg, 2006.

[18] Cesa-Bianchi, Nicolò, Claudio Gentile, and Luca Zaniboni. "Hierarchical classification: combining Bayes with SVM." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.

[19] Agrawal, Rahul, et al. "Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages." *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013.

[20] Clare, Amanda, and Ross D. King. "Knowledge discovery in multi-label phenotype data." *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer Berlin Heidelberg, 2001.

# 10 附录清单

## 10.1 源程序清单（Python 语言，仅列出核心函数）

位点编码：encode_snp.py

| | |
|---|---|
| SNP_encode | 碱基对编码 |

位点关联性分析：correlation_snp.py

| | |
|---|---|
| display_person_distance | 检验个体独立 |
| significance_test | 位点显著性检验 |
| feature_selection | 特征选择模型 |
| snp_validation | 位点相关评估 |

基因关联性分析：correlation_gene.py

| | |
|---|---|
| statistic_candidate_gene | 统计可疑位点所在基因 |
| select_most_probable_gene_regression | 选择最可能基因 |
| gene_validation | 基因相关评估 |

多性状关联性分析：correlation_multiphenos.py

| | |
|---|---|
| statistic_phenos_correlation | 症状间关联性分析 |
| multiphenos_learning | 多症状学习 |
| multiphenos _validation | 多症状位点相关评估 |